Collection of Biostatistics Research Archive
COBRA Preprint Series

Year 2011 *Paper* 81

A Bayesian Model Averaging Approach for Observational Gene Expression Studies

Xi Kathy Zhou* Fei Liu†

Andrew J. Dannenberg‡

*Division of Biostatistics and Epidemiology, Department of Public Health, Weill Cornell Medical College, kaz2004@med.cornell.edu

†Department of Business Analytics and Mathematical Sciences, IBM Watson Research Center

‡Department of Medicine, Weill Cornell Cancer center, Weill Cornell Medical College

http://biostats.bepress.com/cobra/art81

A Bayesian Model Averaging Approach for Observational Gene Expression Studies

Xi Kathy Zhou, Fei Liu, and Andrew J. Dannenberg

Abstract

Identifying differentially expressed (DE) genes associated with a sample characteristic is the primary objective of many microarray studies. As more and more studies are carried out with observational rather than well controlled experimental samples, it becomes important to evaluate and properly control the impact of sample heterogeneity on DE gene finding. Typical methods for identifying DE genes require ranking all the genes according to a pre-selected statistic based on a single model for two or more group comparisons, with or without adjustment for other covariates. Such single model approaches unavoidably result in model misspecification, which can lead to increased error due to bias for some genes and reduced efficiency for the others. We evaluated the impact of model misspecification from such approaches on detecting DE genes and identified parameters that affect the magnitude of impact. To properly control for sample heterogeneity and to provide a flexible and coherent framework for identifying simultaneously DE genes associated with a single or multiple sample characteristics and/or their interactions, we proposed a Bayesian model averaging approach which corrects the model misspecification by averaging over model space formed by all relevant covariates. An empirical approach is suggested for specifying prior model probabilities. We demonstrated through simulated microarray data that this approach resulted in improved performance in DE gene identification compared to the single model approaches. The flexibility of this approach is demonstrated through our analysis of data from two observational microarray studies.

A Bayesian Model Averaging Approach for Observational Gene Expression Studies

XI KATHY ZHOU
Division of Biostatistics and Epidemiology, Department of Public Health
Weill Cornell Medical College
kaz2004@med.cornell.edu

FEI LIU
Department of Business Analytics and Mathematical Sciences
IBM Watson Research Center

ANDREW J. DANNENBERG
Department of Medicine, Weill Cornell Cancer Center
Weill Cornell Medical College

Abstract

Identifying differentially expressed (DE) genes associated with a sample characteristic is the primary objective of many microarray studies. As more and more studies are carried out with observational rather than well controlled experimental samples, it becomes important to evaluate and properly control the impact of sample heterogeneity on DE gene finding. Typical methods for identifying DE genes require ranking all the genes according to a pre-selected statistic based on a single model for two or more group comparisons, with or without adjustment for other covariates. Such single model approaches unavoidably result in model misspecification, which can lead to increased error due to bias for some genes and reduced efficiency for the others. We evaluated the impact of model misspecification from such approaches on detecting DE genes and identified parameters that affect the magnitude of impact. To properly control for sample heterogeneity and to provide a flexible and coherent framework for identifying simultaneously DE genes associated with a single or multiple sample characteristics and/or their interactions, we proposed a Bayesian model averaging approach which corrects the model misspecification by averaging over model space formed by all relevant covariates. An empirical approach is suggested for specifying prior model probabilities. We demonstrated through simulated microarray data that this approach resulted in improved performance in DE gene identification compared to the single model approaches. The flexibility of this approach is demonstrated through our analysis of data from two observational microarray studies.

Keywords: Bayesian model averaging; differential gene expression; microarray; observational study.

1 Introduction

In recent years, as the rapid advances in biotechnology have markedly driven down the cost of microarray experiments, more and more large scale studies are carried out with heterogeneous samples conveniently collected from subjects of different phenotypic characteristics and exposure histories. Such microarray studies are considered observational rather than experimental in nature (Potter, 2003) because the effects of confounding or correlation in covariates need to be properly handled. The sample complexity of such studies presents both opportunities and challenges to the analysis. Considering the differential gene expression studies, with multifaceted sample characteristics, one may explore more complex questions that are not possible with a more homogeneous sample. It should be possible, for example, to identify differentially expressed (DE) genes associated with not just one sample characteristic but multiple characteristics and/or their interactions. However, the existing methodologies are not adequate to address those questions. For example, in a recent study by Boyle et al. (2010) to evaluate the effects of smoking on the transcriptome of human oral mucosa, besides the standard question of identifying genes affected by smoking among all subjects, the investigators were also interested in identifying genes that may be affected by smoking differently between males and females. In another study involving smokers and never smokers (Carolan et al., 2008), micorarray data were obtained for an unbalanced lung airway epithelium sample involving different tissue sites from subjects of different gender, age and ethnicity. An interesting question is to identify DE genes associated with either a single or multiple sample characteristics. Direct application of current available approaches to these questions does not provide a coherent solution and has clear limitations.

Methods for identifying DE genes are typically based on the ranking of statistics for between group differences associated with one sample characteristic (also known as a factor or a covariate), such as the t-, F- statistics, their non-parametric counterparts, their modified forms, or the Bayesian versions (see Jeffery et al. (2006) for a review of the various approaches). These methods are suited for well controlled experiments. Their lack of control for confounding factors attracts increasing concern when applied to observational microarray studies (Potter, 2003; Webb et al., 2007; Troester et al., 2009). Indeed, with observational samples, the results may be confounded by a variety of sample characteristics, such as age, sex, genetic profile, exposure and treatment history, etc. This confounding effect can lead to an increased number of false discoveries. Recent studies by Scheid and Spang (2007) and Leek and Storey (2007) suggested that hidden traces of unknown confounders may exist in DE gene studies and that ranking statistics need to be adjusted accordingly. To account for the effects of possible confounders, several approaches have been adapted from traditional observational studies and applied to microarray data (Smyth, 2004; Hummel et al., 2008). These approaches deal with possible confounding either through multiple regression analysis

that adjusts for known confounders or unrecorded confounders represented by surrogate variables (Leek and Storey, 2007) or through a matched study design (Heller et al., 2009).

Regardless of covariate adjustment, the aforementioned approaches rank the genes based on the effect sizes estimated using the same model for all genes. Such a single model approach can be problematic for high dimensional microarray data because different genes may be involved in different biological processes and their expression may be affected by different sets of covariates. More specifically, such an approach leads to model misspecification for a certain proportion of the genes and does not offer the same level of accuracy and efficiency for the effect size estimation for genes under investigation.

To avoid model misspecification in microarray data analysis, an ideal solution could be to apply different models to different sets of genes whereby each model contains only the set of covariates relevant to the genes it is describing. Yet, identifying appropriate models for different sets of genes can be challenging since model uncertainties make it difficult to identify a single best model. Bayesian model averaging (BMA) approach offers an attractive alternative solution to this problem. Hoeting et al. (1999) provides a review of this approach in more traditional settings. In recent years, BMA approaches have been developed to handle various problems involving high throughput genetic data. For example, Yeung et al. (2005) developed a BMA approach for improved sample classification using microarray data. Xu et al. (2011) and Wu et al. (2010) developed BMA approaches for the gene association studies. Sebastiani et al. (2006) developed a BMA approach for differential gene expression analysis under two different distributional assumptions of the data. Ishwaran and Rao (2003) and Conlon et al. (2006) developed Bayesian model selection/averaging approaches for differential gene expression analysis under ANOVA-based models. All these approaches are computationally expensive as MCMC simulation are needed to obtain estimates of model parameters. In this study, we propose a BMA approach for observational microarray studies based on linear regression models. It does not require MCMC simulations for estimating model parameters and offers a flexible and coherent framework to identify simultaneously DE genes associated with a single factor, multiple factors and/or their interactions.

In the next section, we discuss the limitations of the single model approaches. In particular, we evaluate the impact of model misspecification from such approaches on DE gene finding. We also identify parameters that affect the magnitude of impact. In Section 3, a BMA approach to DE gene finding is proposed. This approach intends to properly control for sample heterogeneity and to account for model uncertainty. In section 4, we compared the performances of ranking statistics based on a simple model, a complex model and the BMA approach in simulated microarray studies. Finally, in section 5, the BMA approach was applied to two existing microarray data sets. Our analysis supports the utility of the BMA method as a

robust and general tool for detecting DE genes in observational microarray studies.

2 Limitation of the Single Model Approaches

In this section, we consider a general framework to describe gene expression variations in microarrays. Under this framework, we argue that the single model approaches to DE gene detection, are overly simplified and subjected to the impact of model misspecification. For example, they can result in the following two forms of model misspecification: (a) the omission of relevant covariates when a simple model is used, and (b) the inclusion of irrelevant covariates when a complex model is used. The consequences of these two forms of model misspecification have been discussed extensively in the linear regression setting (Rao, 1971, 1973; Rosenberg and Levy, 1972). The implication of these results, however, has not been fully investigated in DE gene studies. In this section, we evaluate the consequences of model misspecification from the single model approaches on performance measures often used in DE gene studies, including the false discovery rate (FDR) and sensitivity. We conclude this section with a summary of the main results.

2.1 Notation

We consider an observational microarray study which aims to identify DE genes associated with different values of a factor X_1, for example, cigarette smoking exposure. Expression profiles of J genes are obtained for n subjects with different values of X_1. Without loss of generality, a typical model for identifying X_1 related DE genes can be written as:

$$y_{ij} = \beta_{0j} + \beta_{1j}x_{1i} + \cdots + \beta_{kj}x_{ki} + \eta_{ij} \tag{2.1}$$

or

$$y_{ij} = \alpha_{0j} + \alpha_{1j}x_{1i} + \cdots + \alpha_{kj}x_{ki} + \alpha_{(k+1)j}x_{(k+1)i} + \epsilon_{ij} \tag{2.2}$$

where y_{ij} is the normalized and typically log-transformed expression level of gene j in subject i, x_{1i} is the factor level for X_1 in subject i, $x_{2i}, \ldots, x_{ki}$ are factor levels for other factors, denoted by $X_2, \ldots, X_k$, that affect the expression of all the genes, for example, experimental parameters involved in the microarray experiments, and $x_{(k+1)i}$ is the level of a potential confounding factor , X_{k+1}, for example, gender, age, race, alcohol exposure, etc. And η_{ij}, ϵ_{ij} denote normally distributed random errors.

To identify DE genes related to X_1, p-values based on t-statistic of estimate of either β_{1j} or α_{1j} can be used as the ranking statistics. If model (2.1) is used, the relevant t-statistic for gene j is $t_{M_1,1j} = \hat{\beta}_{1j}/sd(\hat{\beta}_{1j})$, where $\hat{\beta}_{1j}$ is the least square estimate of β_{1j}. If model (2.2) is used, the t-statistic for gene j is

calculated as $t_{M_2,1j} = \hat{\alpha}_{1j}/sd(\hat{\alpha}_{1j})$. It can be shown that the two statistics are related as follows

$$t_{M_1,1j} = \frac{S_{1\cdot 23\ldots k}}{S_{1\cdot 23\ldots k+1}} t_{M_2,1j} + \frac{S^{-2}_{k+1\cdot 1\ldots k} b_{k+1,1} e^T_{k+1} Y_j}{sd(\hat{\beta}_{1j})}. \tag{2.3}$$

where $S^2_{k+1\cdot 1\ldots k}$, b_{k+1} and e_{k+1} are the residual sum of squares, least square parameter estimates and residual, respectively, from the following auxiliary regression equation

$$X_{k+1} = X b_{k+1} + e_{k+1} \tag{2.4}$$

where $X = (X_1, \ldots, X_k)$. And $S^2_{1\cdot 23\ldots k.k+1}$ is the residual sum of squares for the auxiliary regression with X_1 as the outcome and $X_2, \ldots, X_{k+1}$ as the covariates.

For an observational microarray study, such single model approach with or without covariate adjustment has an intrinsic limitation which is that neither model can be the true model for all the genes. For the above discussed hypothetical microarray study, model (2.1) is the true model only for genes not related to X_{k+1} (X_{k+1} null genes, or M_1 genes), and model (2.2) is the true model only for genes related to X_{k+1} (X_{k+1} DE genes, or M_2 genes). Based on these considerations, a multi-model approach that uses p-values of $t_{M_1,1\cdot}$ to rank the M_1 genes and p-values of $t_{M_2,1\cdot}$ to rank the M_2 genes is preferable.

The performance difference between the single model and the multi-model approaches can be compared by utilizing the relationship between the two t-statistics. Let $F_1(t)$ and $F_2(t)$ be the density distributions of the ranking statistics $t_{M_1,1\cdot}$ and $t_{M_2,1\cdot}$, respectively. Under the multi-model approach, the density distribution of the ranking statistics can be written as

$$F(t) = (1-f)F_1(t) + fF_2(t)$$

where f is the proportion of M_2 genes. $F_1(t)$ and $F_2(t)$ can further be written as

$$\begin{aligned} F_1(t) &= (1-p_1)F_{10}(t) + p_1 F_{11}(t) \\ F_2(t) &= (1-p_2)F_{20}(t) + p_2 F_{21}(t). \end{aligned}$$

where p_1 and p_2 are the proportions of DE genes in M_1 and M_2 genes, $F_0(t)$ and $F_1(t)$ are distributions of the test statistic for the null and DE genes, respectively. For a given cut-off $c > 0$, the false discovery rate

and sensitivity can be calculated as

$$FDR(c) = \frac{(1-f)(1-p_1)[1-F_{10}(c)]}{(1-f)[1-F_1(c)]+f[1-F_2(c)]} + \frac{f(1-p_2)[1-F_{20}(c)]}{(1-f)[1-F_1(c)]+f[1-F_2(c)]} \quad (2.5)$$

and

$$S(c) = 2(1-f)p_1[1-F_{11}(c)] + 2fp_2[1-F_{21}(c)].$$

We discuss the impact of the two single model approaches on the FDR and sensitivity separately.

2.2 Single model without covariate adjustment

When Model (2.1) is used, the FDR can be written as

$$FDR^{M_1}(c) = \frac{(1-f)(1-p_1)[1-F_{10}(c)]}{(1-f)[1-F_1(c)]+f[1-F_2^{M_1}(c)]} + \frac{f(1-p_2)[1-F_{20}^{M_1}(c)]}{(1-f)[1-F_1(c)]+f[1-F_2^{M_1}(c)]}$$

The sensitivity can be written as

$$S^{M_1}(c) = 2(1-f)p_1[1-F_{11}(c)] + 2fp_2[1-F_{21}^{M_1}(c)].$$

The superscript M_1 is used to denote that the distribution of t-statistic is derived from model (2.1), which is mis-specified for the M_2 genes because of omitting relevant covariate X_{k+1}.

Omission of relevant covariate leads to bias in the model parameter estimates (Rao, 1971). Specifically, the bias can be written as

$$\text{Bias}(\hat{\beta}_{1j}) = E(S^{-2}_{k+1\cdot 1\ldots k} b_{k+1,1} e_{k+1}^T Y_j) = \alpha_{k+1,j} \cdot b_{k+1,1}, \quad (2.6)$$

where $b_{k+1,1.23\ldots k}$ is the least square estimate of the parameter associated with X_1 in the auxiliary regression (2.4). Therefore, we have for the M_2 gene j

$$E(t_{M_1,1j}) \approx \frac{S_{1\cdot 23\ldots k}}{S_{1\cdot 23\ldots k+1}} \left[E(t_{M_2,1j}) + \frac{b_{k+1,1}\alpha_{k+1}}{\sigma_{2j}/S_{1\cdot 23\ldots k+1}} \right].$$

It is also known that $S^2_{1\cdot 23\ldots k.k+1} \leq S^2_{1\cdot 23\ldots k}$.

Therefore, for the M_2 DE genes, because $t_{M_1,1j}$ can be greater or less than $t_{M2,1j}$ depending on the

values of α_1 and $Bias(\hat{\beta}_1)$, $F_{21}^{M_1}(t)$ is unlikely to be systematically different from $F_{21}(t)$ and results in great changes in sensitivity.

However, for the M_2 null genes, the above results indicate $E|t_{M_1,1j}| \geq E|t_{M_2,1j}|$, i.e. the distribution of $t_{M_2,1j}$ for the M_2 null genes moves away from zero. Hence, $1 - F_{20}^{M_1}(c) \geq 1 - F_{20}(c)$. Let a and b be the denominator and numerator of $FDR(c)$ as written in (2.5), respectively. Let δ be the difference between the numerators of $FDR^{M_1}(c)$ and $FDR(c)$, i.e.,

$$\delta = f(1-p_2)\{[1 - F_{20}^{M_1}(c)] - [1 - F_{20}(c)]\},$$

and δ' be the difference between the denominators of the two FDRs

$$\begin{aligned} \delta' = \quad & f(1-p_2)\{[1 - F_{20}^{M_1}(c)] - [1 - F_{20}(c)]\} + \\ & fp_2\{[1 - F_{21}^{M_1}(c)] - [1 - F_{21}(c)]\} \end{aligned}.$$

As discussed above, $[1 - F_{21}^{M_1}(c)]$ is comparable to $[1 - F_{21}(c)]$ because the bias is unlikely to lead to systematic difference between $F_{21}^{M_1}(t)$ and $F_{21}(t)$. Additionally, p_2 generally is much smaller than $1 - p_2$ in microarrays. Therefore, $\delta' \approx \delta$ and $FDR^{M_1}(c)$ can be approximated by $(b+\delta)/(a+\delta)$. Since $(b+\delta)/(a+\delta) \geq b/a$ for any $a > b > 0$ and $\delta \geq 0$, this indicates $FDR^{M_1}(c) \geq FDR(c)$, i.e., increased FDR with this single model approach.

2.3 Single model with covariate adjustment

When Model (2.2) is used, the false discovery rate and sensitivity at a given cut-off can be written as

$$FDR^{M_2}(c) = \frac{(1-f)(1-p_1)[1-F_{10}^{M_2}(c)]}{(1-f)[1-F_1^{M_2}(c)]+f[1-F_2(c)]} + \frac{f(1-p_2)[1-F_{20}(c)]}{(1-f)[1-F_1^{M_2}(c)]+f[1-F_2(c)]}$$

and

$$S^{M_2}(c) = 2(1-f)p_1[1 - F_{11}^{M_2}(c)] + 2fp_2[1 - F_{21}(c)]$$

due to the potential change in the distributions of test statistic for the M_1 genes. And the relationship of the two t-statistics can be written as

$$t_{M_2,1j} = \frac{S_{1\cdot 23\ldots k+1}}{S_{1\cdot 23\ldots k}} t_{M_1,1j} + \frac{S_{k+1\cdot 1\ldots k}^{-2} b_{k+1,1} e_{k+1}^T Y_j}{sd(\hat{\alpha}_{1j})}.$$

It is known that, with the inclusion of an irrelevant covariate, model 2.2 does not result in biased parameter estimate for the M_1 genes. However, since $sd(\hat{\beta}_{1j}) \leq sd(\hat{\alpha}_{1j})$ in general, $E(|t_{M_1,1}|) \geq E(|t_{M_2,1}|)$ for M_1 DE genes. Therefore the distribution $F_{11}^{M_2}(t)$ moves toward 0 and results in $S^{M_2}(c) \leq S(c)$, i.e., reduced sensitivity in detecting DE genes in M_1 genes. As $|t_{M_1,1}|$ in general is likely to be greater than $|t_{M_2,1}|$, F_{10}^{M2} also shrinks toward 0. It is likely that $FDR^{M_2}(c)$ will be comparable to $FDR^{M_1}(c)$. Hence, reduced sensitivity in detecting DE genes in M_1 genes will be the main consequence resulted from applying the complex model for all the genes.

2.4 Summary

The above results suggested that the single model approaches with or without covariate adjustment can lead to inferior performance. It is expected that the impact on FDR and sensitivity could be greater if more X_{k+1}-like covariates exist in the sample. These results will be further demonstrated in the simulation study. The above discussion also suggested that the performance for DE gene detection can be improved by applying the correct model for the right sets of genes. Yet, such knowledge is commonly not available beforehand. In the following section, we propose a BMA approach as a practical substitute for the multi-model approach for DE gene detection that takes into account both sample heterogeneity and model uncertainty.

3 A Bayesian Model Averaging Approach

In this section, we discuss an efficient Bayesian model averaging approach to identifying DE genes associated with a covariate of interest. We consider a series of possible models for describing the expression pattern of each gene. Let $\gamma = (\gamma_1, \ldots, \gamma_K)$, be a binary vector of length K, with each element indicating the inclusion status of the kth covariate in the model,

$$\gamma_k = \begin{cases} 0 & if \beta_k = 0 \\ 1 & if \beta_k \neq 0 \end{cases}$$

Each model in the model space can then be labeled by γ, namely $\mathcal{M}_\gamma$. For gene j, $j = 1, \ldots, J$, the model can be written as

$$\boldsymbol{\mathcal{M}_{\gamma j}}: \quad \boldsymbol{Y_j = \alpha_{\gamma j} 1_n + X_\gamma \beta_{\gamma j} + \mathrm{N}(0, \phi_{\gamma j}^{-1} I_n)},$$

where $\alpha_{\gamma j}$ is the intercept term; $\boldsymbol{X_\gamma}$ is the sub-matrix of X consisting of columns associated with non-zero γ_k; $\boldsymbol{\beta_{\gamma j}}$ and $\phi_{\gamma j}$ are parameters under this model.

The marginal posterior inclusion probability for variable X_k and gene j, is then defined as

$$\boldsymbol{P_{kj} = P(\gamma_{kj} \neq 0|Y_j) = \sum_{\gamma} 1_{\gamma_{kj}=1} \times P(\mathcal{M}_{\gamma j}|Y_j),} \tag{3.1}$$

which is the sum of posterior probabilities of all models that include the covariate of interest. It quantifies strength of association between covariate X_k and the expression level of the jth gene and can be used to rank the DE genes.

The posterior model probability for $\mathcal{M}_{\gamma j}$ can be calculated based on Bayes factors of pairs of models, for example,

$$\boldsymbol{P(\mathcal{M}_{\gamma j}|Y_j) = \frac{p(\mathcal{M}_{\gamma j})BF(\mathcal{M}_{\gamma j} : \mathcal{M}_{0j})}{\sum_{\gamma'} p(\mathcal{M}_{\gamma' j})BF(\mathcal{M}_{\gamma' j} : \mathcal{M}_{0j})}.} \tag{3.2}$$

where $p(\mathcal{M}_{\gamma j})$ is the prior model probability for genes measured in the microarray experiment and the Bayes factor $BF(\mathcal{M}_{\gamma j} : \mathcal{M}_{0j})$ is defined as

$$BF(\mathcal{M}_{\gamma j} : \mathcal{M}_{0j}) = \frac{f(Y_j|\mathcal{M}_{\gamma j})}{f(Y_j|\mathcal{M}_{0j})}$$

i.e., the ratio of marginal likelihood under $\mathcal{M}_{\gamma j}$ and the base model, $\mathcal{M}_{0j}$. Here the null model (i.e., the model with only the intercept term) is used as the base model. For $\mathcal{M}_{\gamma j}$, the marginal likelihood is obtained by integrating out the model parameters from the joint posterior probability

$$\boldsymbol{f(Y_j|\mathcal{M}_{\gamma j}) = \int f(Y_j|\Theta_{\gamma j})\pi(\Theta_{\gamma j})\mathrm{d}\Theta_{\gamma j}}$$

where $\boldsymbol{\Theta_{\gamma j} = (\alpha_{\gamma j}, \beta_{\gamma j}, \phi_{\gamma j})}$, and $\pi(\Theta_{\gamma j})$ is the prior of model parameters.

There are various choices for $\pi(\Theta_{\gamma j})$. To be able to determine the Bayes factor, proper priors are needed. In our study, we utilized the Zellner-Siow prior for model parameters (Zellner and Siow, 1980). Liang et al. (2008) have shown that this prior resolves several consistency issues associated with fixed g-priors while retaining several attractive properties such as adaptivity, good shrinkage properties, robustness and fast marginal likelihood calculation. When comparing two nested models as in our case, the Zellner-Siow prior places a flat prior on common coefficients, i.e., ($\alpha_{\gamma j}$, $\phi_{\gamma j}$), where $\pi(\alpha_{\gamma j}, \phi_{\gamma j}) \propto 1/\phi_{\gamma j}$. And a Cauchy prior on the remaining parameters, i.e., $\boldsymbol{\beta_{\gamma j}}$. The multivariate Cauchy prior can then be represented as a mixture of g-priors with an Inv-gamma(1/2, n/2) prior on g, i.e.

$$\pi(\beta_{\gamma j}|\phi_{\gamma j}) \propto \int N\left(\beta_{\gamma j}|0, \frac{g}{\phi_{\gamma j}}(X_\gamma^T X_\gamma)^{-1}\right)\pi(g)dg,$$

with

$$\pi(g) = \frac{\sqrt{n/2}}{\Gamma(1/2)} g^{-3/2} e^{-n/(2g)}.$$

And the Bayes factor in equation (3.2) can be written in closed form as

$$\begin{aligned} BF(\mathcal{M}_{\gamma j} : \mathcal{M}_{\mathbf{0}j}) &= \int_0^\infty (1+g)^{(n-1-\rho_{\gamma j})/2} \times \\ &\quad [1 + (1 - R^2_{\gamma j})g]^{-(n-1)/2} \pi(g) \mathrm{d}g \end{aligned}$$

where $\rho_{\gamma j}$ denotes the number of covariates included in $\mathcal{M}_{\gamma j}$ and $R^2_{\gamma j}$ is the ordinary coefficient of determination of this model. This quantity can be obtained through direct numerical integration or through the Laplace approximation.

Another step for calculating the posterior model probability, as specified in equation (3.2), is to specify appropriate prior model probabilities. Typically, these prior model probabilities reflect our prior belief about the distribution of the models among the genes in the transcriptome. However, it is often difficult to provide reasonable quantification of the prior belief. One may be tempted to use a uniform prior with the underlying assumption that each of the models is equally likely to be the true model when prior knowledge about the particular gene or the transcriptome is lacking. Yet this assumption can be problematic when thousands of genes are evaluated simultaneously because it puts an unrealistically low weight to the null model. When the resulting posterior model probabilities are used to estimate the posterior expected FDR (*pe*FDR) (Newton et al., 2004), great underestimation can occur (Sartor et al., 2006; Cao et al., 2009). Correctly estimating FDR under the Bayesian framework remains an active research field (Efron, 2008). Recent discussions and attempts have largely been focused on statistics derived from single model approaches (Müller et al., 2007; Cao and Zhang, 2010). In our case, proper control for multiplicity derived from multiple genes and multiple models becomes even more challenging.

We believe that the prior should lead to reasonably well calibrated posterior model probability that measures the model's ability for describing the data. We propose an empirical approach to obtain estimates for the prior model probabilities, $p(\mathcal{M}_{\gamma j})$, under the assumption that the prior probabilities of a given model is the same across genes, i.e., $p(\mathcal{M}_{\gamma j}) = p(\mathcal{M}_\gamma)$. We first estimate the proportion of DE genes described by a non-null model γ, ω_γ, using Bayes factors. Since $BF(\mathcal{M}_\gamma : \mathcal{M}_0) > c$, $c \geq 1$ suggests evidence against the null model (Kass and Raftery, 1995), we can estimate ω_γ as follows

$$\omega_\gamma = \frac{1}{J} \sum_j \mathbf{1}_{[BF(\mathcal{M}_{\gamma j}:\mathcal{M}_{0j}) = max(BF_j)]} \cdot \mathbf{1}_{[BF(\mathcal{M}_{\gamma j}:\mathcal{M}_{0j}) > c]},$$

where BF_j is a vector of null based Bayes factors for gene j. Therefore, ω_γ represents the proportion of genes for which model γ is the best model in terms of Bayes factors. Given that Bayes factors based on the Zellner-Siow prior is consistent for model selection whether or not the true model is null (Liang et al., 2008), this estimator is a consistent estimator of the proportion of genes expressing in a pattern specified by the model. In our simulation studies, we found that fixing c at 1 resulted in ω_γ being close to the truth in most settings. Second, we argue that if the prior model probabilities, p_γ, results in the equality between the overall peFDR under $\mathcal{M}_\gamma$ and $1-\omega_\gamma$, reasonable calibration of the posterior model probabilities can be achieved. Therefore, the prior model probabilities, $p(\mathcal{M}_\gamma)$, can be derived from the following relationship

$$\omega_\gamma = \frac{1}{J}\sum_j \frac{BF(\mathcal{M}_{\gamma j}:\mathcal{M}_{0j})p(\mathcal{M}_\gamma)}{\sum_{\gamma'} BF(\mathcal{M}_{\gamma' j}:\mathcal{M}_{0j})p(\mathcal{M}_{\gamma'})},$$

using an iterative procedure under the constraint $\sum_\gamma p(\mathcal{M}_\gamma)=1$. At present stage, theoretical justification for this prior choice for multiplicity control is still lacking. We resort to the simulation study to show that this prior choice led to improved performance in both the ranking of the genes and in direct FDR estimation compared with the uniform prior.

4 Simulation Study

There were two reasons to carry out the simulation study. First, we intended to demonstrate the performance difference in DE gene detection between the single model approaches with or without covariate adjustment, and the "gold standard" multi-model approach where covariate adjustments were applied to appropriate genes. The interplay between the bias and efficiency as the source for the difference will be explored. Secondly, the performance of the BMA approach as a practical substitute for the multi-model approach will be evaluated. Sensitivity to the choice of prior model probabilities will be discussed.

4.1 Simulation of microarray data

The microarray data were simulated to mimic an observational study for identifying genes associated with a binomial factor, for example, the smoking status (s), in a sample with two confounders, for example, gender (g) and heavy alcohol drinking (d). Each microarray data set consists of the expression of 10000 genes from n subjects, in which half were heavy smokers and half were never-smokers. We limited our attention to the imbalanced data where the model with adjustment to g and d could presumably help. We assumed that there were 25% females and 75% heavy drinkers in smokers, and 75% females and 25% heavy drinkers in nonsmokers. Additionally, in smoking males, 87% are assumed to be heavy drinkers, whereas in smoking

females, 60% are assumed to be heavy drinkers; in non-smoking males, the proportions were 40% and 13%, respectively. Therefore, the data show a complex pattern of correlation, i.e., not only are gender and drinking status correlated with smoking status individually, gender is also correlated with smoking status. Gene expression for each subject was simulated based on the following model

$$y_{ij} = \beta_{1j} s_i + \beta_{2j} g_i + \beta_{3j} d_i + \varepsilon_{ij}$$

where $\beta_{.j}$ takes either 0 or non-zero values generated from normal distributions with variances generated following procedures similar to that described by Smyth (2004). Detailed procedures for generating the simulated microarray data are provided in the supplemental material. Each simulation setting was characterized by values of the following parameters: f_s, f_g, and f_d, the proportion of genes affected by smoking (s), gender (g), or heavy drinking (d), respectively, and n, the sample size. Both moderate and relatively large sample sizes were considered, $n = 40$ and $n = 80$. For each setting, we simulated 10 microarray data sets. The reported results were averaged over the results obtained for each data set.

4.2 Performance of the single model approaches

In this section, we show the performance difference between the single model approach and the gold standard multi-model approach. To identify DE genes associated with s, p-values for the effects of s measured through four approaches were used as ranking statistics. Specifically, we used the single model approach without covariate adjustment (SM_1), the single model approach with covariate adjustment (SM_2), the surrogate variable analysis (SVA) method developed by Leek and Storey (2007) and the gold standard multi-model approach (MM) where the expression data of the DE genes were fit with their respective true models. Under the SVA approach, we assumed that information regarding gender and drinking status was not recorded and surrogate variables were generated in place of them. Under the MM approach, the adjustment for g and/or d is applied only to genes truly affected by g and/or d. Thus the results from the MM approach can be viewed as the gold standard. The sensitivity and FDR corresponding to each ranking statistic were obtained. To show the interplay of bias and efficiency on these performance measures, we also quantified the contribution to these measures from genes not associated with g and d, denoted as $g0d0$ genes.

Table 1 shows the performance difference between the single and multi- model approaches among top ranked genes identified with a p-value cut-off of 0.001. We can see that, as discussed in Section 2, SM_1 led to large increase in total FDR compared to MM. The magnitude of difference increased with the sample

Table 1: False discovery rate (FDR) and sensitivity (S), in %, among the top smoking related genes identified with a p-value cut-off of 0.001 using ranking statistics based on the single model approach without covariate adjustment (SM_1), the single model approach with covariate adjustment (SM_2), the surrogate variable analysis approach (SVA), and the "gold standard" multi-model approach (MM). FDR and sensitivity arising from $g0d0$ genes (i.e. genes not associated with d and g) were included. Microarray datasets were simulated based on various settings defined by proportion of genes associated with each covariate: f_s, f_g, f_d, and n.

Methods	FDR_{g0d0}	FDR_{total}	S_{g0d0}	S_{total}	FDR_{g0d0}	FDR_{total}	S_{g0d0}	S_{total}
	n=40				n=80			
$f_s = 0.10$, $f_g = 0.05$, $f_d = 0$								
SM_1	4.2	6.5	14.1	14.9	2.3	8.2	28.5	30.2
SM_2	6.1	6.5	10.0	10.4	2.2	2.5	23.1	24.3
SVA	6.2	6.7	9.3	9.7	2.2	2.3	22.7	24.0
MM	4.4	4.8	14.1	14.5	2.4	2.6	28.5	29.7
$f_s = 0.05$, $f_g = 0.10$, $f_d = 0$								
SM_1	8.5	18.0	12.9	14.5	3.6	22.9	26.6	29.8
SM_2	10.5	11.7	9.4	10.6	5.1	5.7	21.0	23.4
SVA	10.3	11.6	9.0	10.2	5.4	6.2	20.7	23.1
MM	9.5	9.5	12.9	14.1	4.5	5.0	26.6	29.0
$f_s = 0.1$, $f_g = 0.05$, $f_d = 0.05$								
SM_1	4.1	8.5	13.4	15.0	2.6	12.8	26.4	29.5
SM_2	6.8	7.4	8.4	9.3	3.0	3.2	19.9	22.1
SVA	7.1	8.0	8.2	9.1	2.9	3.1	19.3	21.5
MM	4.4	4.9	13.4	14.3	3.0	3.1	26.4	28.8
$f_s = 0.05$, $f_g = 0.10$, $f_d = 0.10$								
SM_1	4.7	19.4	12.6	15.7	3.4	36.9	25.1	31.3
SM_2	9.9	12.6	7.9	9.6	4.5	6.1	18.4	22.6
SVA	10.7	13.1	7.7	9.4	4.6	6.0	18.0	22.4
MM	5.9	8.2	12.6	14.4	5.4	6.5	25.1	29.6

size, the proportion of the genes associated with the confounder and the number of the confounders. On the other hand, the difference in FDR contributed from the $g0d0$ genes remained small. Hence, the results suggested that bias in effect estimation among genes associated with the confounders was the main source for the FDR increase. SM_2 and SVA showed slightly greater FDR compared to MM. This increase came mainly from $g0d0$ genes and suggested that the effects of the efficiency loss could have a negative impact on the total FDR, particularly in small sample size settings. A more notable limitation of SM_2 and SVA was the loss of sensitivity. Compared to MM, the magnitude of sensitivity loss increased slightly with sample size and the number of confounders.

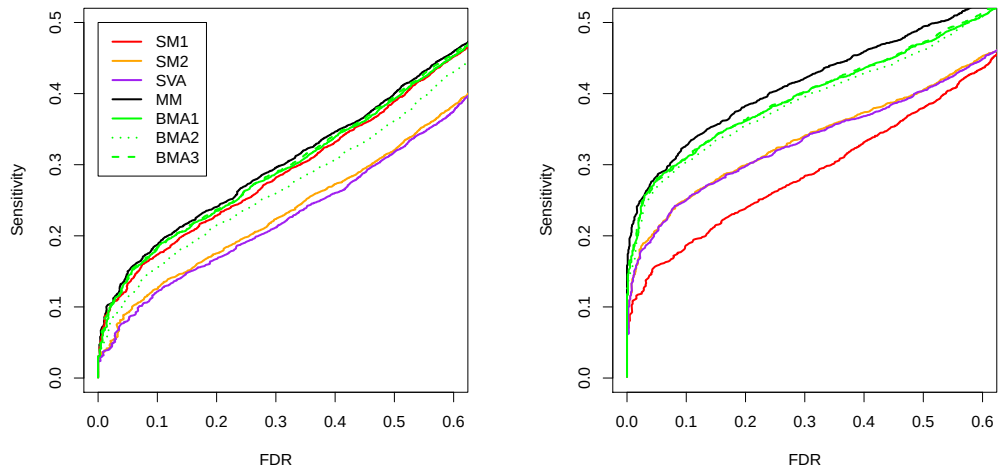
Table 2: Power of different methods for identifying genes differentially expressed between smokers and non-smokers at 5% FDR under different simulation settings.

f_s	f_g	f_d	SM_1	SM_2	SVA	BMA_1	BMA_2	BMA_3	MM
$n = 40$									
0.10	0.05	0	139	96	83	145	119	149	155
0.10	0.05	0.05	126	80	72	137	124	137	150
0.05	0.10	0	42	31	31	51	46	52	52
0.05	0.10	0.10	46	30	26	56	49	57	58
$n = 80$									
0.10	0.05	0	286	294	290	346	335	344	356
0.10	0.05	0.05	239	250	248	317	308	318	334
0.05	0.10	0	94	113	110	147	135	146	152
0.05	0.10	0.10	82	108	106	145	138	142	147

4.3 Performance of the BMA approach

In this section, we examine the performance of the proposed BMA approach in comparison with the single model and the gold standard multi-model approaches. To evaluate the effects of prior choice on the performance of the BMA approach, we considered three prior model probability choices: the proposed empirical prior obtained using the two step approach (BMA_1), the uniform prior (BMA_2), and the true proportion of genes for each model (BMA_3). The posterior inclusion probability of s was used as the ranking statistics. The number of genes identified by each methods at 5% FDR were compared in Table 2. We can see that the BMA approaches had greater power in detecting DE genes compared to the SM approaches in general and the performance came close to that of the MM approach. In fact, in all the simulated settings, the BMA approaches, particularly BMA_1 and BMA_3, showed sensitivity close to the MM approach for a given FDR threshold and greater than the single model approaches. Fig 1 showed the magnitude of performance difference in two representative settings. The BMA approaches appeared to be relatively insensitive to the choice of prior model probabilities for gene ranking. In the supplemental material, we showed additional results that suggest that the ranking performance of BMA_1 is relatively robust to the choice of c and to the misspecification of the model space.

Besides providing proper ranking of the gene, it is often useful to estimate the FDR of the finding and quantifying the proportion of DE genes in the transcriptome. Therefore, we also evaluated how well the FDR could be estimated based on the ranking statistics. For the p-value based approach, FDR and the proportion of DE genes were estimated using the approach by Storey et al.(Storey, 2002; Storey and Tibshirani, 2003). For the Bayesian model averaging approach, the *pe*FDR was directly estimated based on



(a). $f_s = 0.1$, $f_g = 0.05$, $f_d = 0$, $n = 40$. (b). $f_s = 0.05$, $f_g = 0.1$, $f_d = 0.1$, $n = 80$.

Figure 1: Sensitivity vs. FDR curves in two simulation settings.

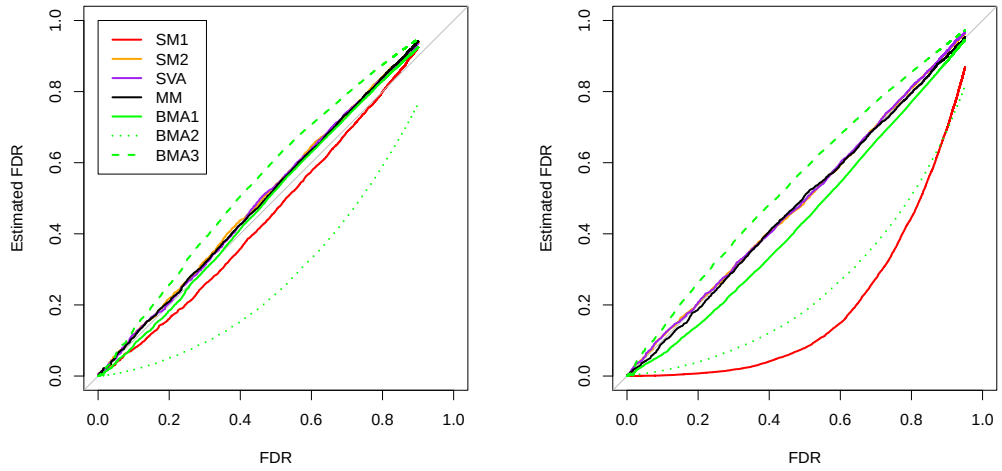
the posterior inclusion probability (Newton et al., 2004), i.e.

$$peFDR_k(p) = \sum_j (1 - P_{kj}) \cdot \mathbf{1}_{[P_{kj} \le p]} / \sum_j \mathbf{1}_{[P_{kj} \le p]},$$

where $0 < p \le 1$ and P_{kj} is the posterior inclusion probability of variable k for gene j. Fig 2 shows the estimated FDR vs. the true FDR in two representative settings. We can see that using p-values from SM_1 in studies with confounder associated genes, the estimated FDR was smaller than the true FDR. The magnitude of underestimation increased with the sample size and the proportion of the confounder associated genes. On the other hand, the FDR estimated using p-values from SM_2 or MM was very close to the true FDR. The accuracy of the peFDR, as observed by other researchers, appeared to be sensitive to the prior choice. peFDR obtained based on BMA_3, the Bayesian model averaging approach with uniform prior can greatly underestimate the FDR. peFDR obtained based on BMA_1 showed improved accuracy in FDR estimation. The results from our simulation also suggest that the peFDR based on BMA_1 are close to true FDR in all simulated settings. BMA_3 appeared to result in peFDR that slightly overestimated the FDR. Level of sensitivity of the BMA_1 approach to the choice of c and model space misspecification can be found in the supplemental material.

5 Application to the Observational Micorarray Datasets

We applied the BMA approach to two observational microarray studies involving healthy current smokers and never smokers. Through the application, we intended to demonstrate the complex relationship between



(a). $f_s = 0.1$, $f_g = 0.05$, $f_d = 0$, $n = 40$. (b). $f_s = 0.05$, $f_g = 0.1$, $f_d = 0.1$, $n = 80$.

Figure 2: Estimated FDR vs. true FDR in two simulation settings.

the gene expression pattern and sample characteristics and the flexibility of the Bayesian model averaging approach in capturing and quantifying such relation in a unified and coherent framework.

5.1 Microarray study of airway epithelium samples

The first dataset we used came from a microarray study (GSE10006) of a total of 87 current and never smokers (Carolan et al., 2008). The microarray analyses were carried out on airway epithelium samples obtained from these subjects. The data were preprocessed with Affymetrix MAS method. After excluding genes whose expression was deemed absent or marginal among all subjects, the data we used consisted of expression profiles of 44085 genes from the Affymetrix HGU133plus2 chip for each subject. Besides smoking status, information on age, gender, race and site of the tissue was available. We limited our analyses to the data from 60 healthy subjects. Individuals with known lung disease were excluded. The samples were heavily unbalanced, the proportion of smokers was greater in female participants than in males (86% vs. 57%), the proportion of large airway samples was slightly larger in females than in males (57% vs. 46%), and the proportion of caucasian participants was larger in females compared to males (43% vs. 37%).

Since the subject characteristics were captured in five covariates, a total of 2^5 models were included in the model space. Interaction terms were ignored. The BMA approach allowed for simultaneous assessment of the association between the gene expression and each of the sample characteristics. And it allows for straightforward estimation of both the total proportion of the DE genes in the transcriptome and the proportion of DE genes associated with each covariate based on Bayes factors. The application showed a complex picture of the expression pattern in the epithelium microarray study. A total of 52% of the genes were estimated to be differentially expressed. The estimated proportions of DE genes for association with $smoking$,

$site$, $gender$, $race$, and age were 15%, 24%, 5%, 5% and 3%, respectively. By controlling the *pe*FDR at 5%, we identified a number of DE genes associated with $smoking$ (1742), $site$ (5019), $gender$ (49), $race$ (33) and age (5). The complex expression patterns were illustrated in Fig 3 where we show the expression pattern of the top 20 genes associated with $smoking$, $gender$, $site$ and $race$, respectively.

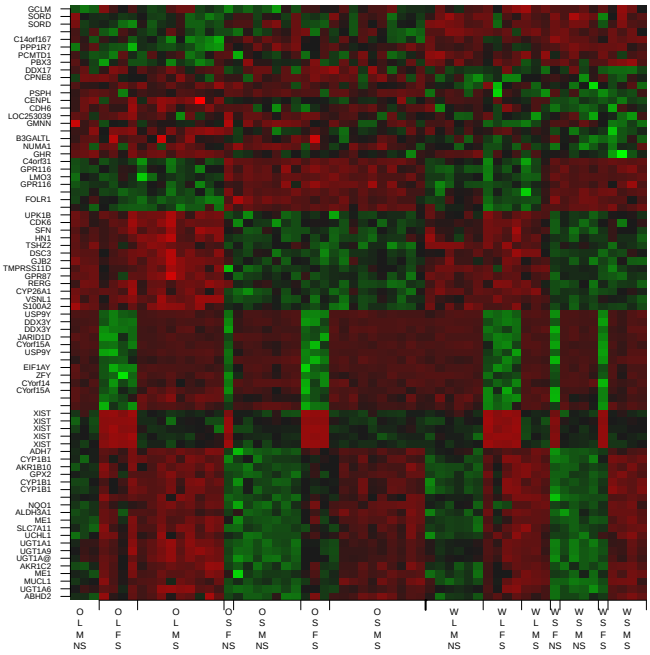


Figure 3: Gene expression intensities for the top 20 genes associated with each of the four covariates ($smoking$, $gender$, $site$, and $race$) identified by using BMA_1. Labels along the x-axis show the characteristics of a sample subgroup. From top to bottom, the label represents categories of $race$ (Others vs. White; O vs. W), $site$ (Large airway vs. Small airway; L vs. S), $gender$ (Male vs. Female; M vs. F), and $smoking$ status (Non-Smoker vs. Smoker; NS vs. S). For example "O L M NS" indicates the subgroup with the following characteristics: Other races (i.e., non-white), Large airway sample, Male, Non-Smoker.

The results also revealed complex roles of some of these DE genes which showed strong association with multiple sample characteristics. For example, among the top 1742 $smoking$ related DE genes, 790, 40 and 9 probesets of genes were also identified as hits for association with tissue $site$, $gender$ and $race$, respectively. Additionally, there were 26 genes identified as hits for association with three or more sample characteristics, mostly $smoking$, $site$ and $gender$. The BMA approach allows for assessing jointly genes' association with multiple sample characteristics. For example, the joint posterior inclusion probability of $smoking$, $site$, and $gender$ can be obtained by summing over the posterior probabilities of models containing all three covariates. *pe*FDR can then be derived similarly using this posterior inclusion probability. The analysis identified 4 genes, IRX2, TMEM17, UGT1A3, and NRARP as hits for joint association with the three characteristics at 5% peFDR. The existence of such genes suggested a connection between tobacco smoking

and the functions of these genes which were partly revealed through their association with the phenotype of the subjects from whom samples were obtained. Results from such analysis offers additional important information that is useful for generating new hypotheses and insights into the effects of tobacco smoke on the transcriptome.

As discussed in the previous sections, given the existence of genes associated with various sample characteristics, single model approaches were subjected to the effects of increased bias or reduced power in unbalanced study design. For the epithelium microarray data, we saw large differences in gene rankings derived from the Bayesian model averaging approach and the single model approaches, including the single model approach with $smoking$ status as the only covariate ($SM1$), the single model approach that included all recorded covariates ($SM2$), and the surrogate variable approach (SVA). Among the top 1000 smoking related DE genes identified by each method, the agreement was merely 17% among all four methods. Specifically, the SVA approach produced gene lists that were vastly different from the gene lists produced by the other approaches, where more than half of the top 1000 genes had ranks beyond 1000 by the other three methods (see the Venn diagram in supplemental material). Careful examination of the gene lists produced by the SVA approach suggested possible effects of over-fitting as the SVA approach adjusted for a total of 12 surrogate variables for each gene. The agreement was about 62% for the $SM1$, $SM2$, and $BMA1$ approaches, i.e., 62% were ranked within top 1000 by all three methods. And the agreement between BMA1 and each of the single model approaches was less than 80%. These differences were driven by the genes whose expression patterns were not adequately captured by the single model.

5.2 Microarray study of oral mucosa samples

The second dataset we used included a total of 79 age and gender matched healthy smokers and never smokers. The microarray analyses were carried out on oral mucosa samples obtained from these subjects through buccal biopsies. The preprocessed microarray data consisted of 24103 probesets of genes from the Affymetrix HGU133plus2 chip for each subject. Information regarding age, gender, smoking status were available. Details of the original study were provided in Boyle et al. (2010).

The study samples were balanced in terms of gender between smokers and non-smokers. Therefore, single model approaches with or without adjustment for gender would provide similar results. However, one interesting biological question was whether there were genes affected by smoking differently between the males and females. In this context, direct application of the single model approach could lead to confusing results. For example, at 5% estimated FDR, the single model without adjustment for the interaction term resulted in 1254 hits for association with $smoking$, while the model adjusted for both $gender$ and $gender \times$

$smoking$ interaction led to the identification of only 2 genes as hits for association with $smoking$ and no genes were identified as hits for $smoking \times gender$ interaction. Such large difference in DE gene assessment between different models is difficult to reconcile and interpret under the single model framework. Yet, such difference can be expected if there are genes associated with the interaction because the two variables, $smoking$ and $smoking \times gender$ interaction, are correlated. Joint testing of the effects of $smoking$ and $smoking \times gender$ interaction led to the identification of 345 DE genes with the likelihood ratio test. However, this method can not quantify the relative contribution from the two variables. We therefore applied the BMA approach to these data to illustrate the flexibility and usefulness of this approach to handle possible interaction effects.

In this application, the model space consists of a total of 16 models including the null model, three models with $smoking$ and/or $gender$ as main effects only and 12 models for different patterns that could arise from interaction between $smoking$ and $gender$. For the oral mucosa data, our analysis estimated that about 21% of the probesets are differentially expressed, in which, about 11%, 1.6%, and 9% were associated with $smoking$, $gender$ and $smoking \times gender$ interaction, respectively. Controlling the *pe*FDR at 5%, our approach identified a total of 595 probesets as hits associated with smoking through either the main effect, the interaction effect or both. Specifically, 291 of these genes were associated with $smoking$ primarily through the main effect, 10 were associated with $smoking$ primarily through the interaction effect, while for the rest of these genes various degrees of association were contributed from the interaction term.

By comparing the $smoking$ related DE genes identified by the single model approaches and the BMA approach, we noted that the difference was mainly from genes that were over/under expressed in only one subgroup of the subjects, female smokers. Neither the model with $smoking$ status as the only covariate nor the full model adjusted for both the $gender$ and the $smoking \times gender$ interaction were able to adequately capture the strength of association for this group of genes and properly rank them due to either increased bias or decreased power. Table 3 showed the posterior inclusion probabilities and ranks based on different approach for a few of these genes. Large difference in the rankings by different methods can be seen.

6 Discussion

In the past decade, microarray technology has greatly increased our ability to simultaneously interrogate the expression of tens of thousands of genes. Use of this technology has contributed to an improved understanding of the molecular basis of various diseases. As one of the primary tools for such studies, methods for finding DE genes have also been refined over time. Various approaches have been proposed to deal with multiple issues in microarray data. Yet, from the modeling perspective, many approaches have ignored

Table 3: Posterior covariate inclusion probabilities of genes identified as hits for association with s, g, and $s \times g$ interaction separately, and s or $s \times g$ interaction jointly (denoted as P_s,P_g, $P_{s\times g}$, and $P_{s|s\times g}$, respectively) obtained using BMA_1. Ranks of these genes based on strength of association with these covariates separately or jointly under different methods ($\mathcal{R}^{method}_{covariate/s}$) were also shown.

GSymbol	ProbeSet	P_s	P_g	$P_{s\times g}$	$P_{s\|s\times g}$	$\mathcal{R}_s^{SM_1}$	$\mathcal{R}_s^{SM_2}$	$\mathcal{R}_s^{BMA_1}$	$\mathcal{R}_{s\times g}^{SM_2}$	$\mathcal{R}_{s\times g}^{BMA_1}$	$\mathcal{R}_{s\|s\times g}^{SM_2}$	$\mathcal{R}_{s\|s\times g}^{BMA_1}$
PAFAH2	205233_s_at	0.009	0.001	0.986	0.991	2301	23739	24103	23	1	326	122
CEACAM7	211848_s_at	0.030	0.003	0.974	0.997	518	7436	16401	220	2	169	72
CEACAM7	206199_at	0.042	0.003	0.966	0.998	271	5485	11556	151	3	85	51
COPS7B	225696_at	0.009	0.009	0.963	0.968	4648	23892	24102	164	4	1316	241
PRKX	204060_s_at	0.019	0.040	0.961	0.977	2345	10494	24059	911	5	840	206
CEACAM7	206198_s_at	0.062	0.003	0.944	0.996	336	6082	8367	194	6	111	76
THYN1	218491_s_at	0.012	0.014	0.938	0.946	5113	23484	24096	270	7	1678	335
CD177	219669_at	0.034	0.006	0.931	0.960	1907	12463	14423	583	8	814	271
BACE2	222446_s_at	0.044	0.003	0.929	0.973	1234	15980	11127	83	9	307	218
MARK1	221047_s_at	0.030	0.002	0.927	0.956	2013	23223	16058	11	10	279	287

sample heterogeneity, its impact on the analysis results, and the great opportunity it presents. Since Potter (2003) discussed the need for controlling bias and confounding in observational microarray studies, it has been increasingly recognized that the lack of control for sample heterogeneity could be a barrier to the reproducibility of the study findings. In two editorials (Webb et al., 2007; Troester et al., 2009), improved data analysis methods and better study design have been considered crucial for advancing the field of cancer epidemiology with microarray technology. In particular, Troester et al. (2009) discussed the potential of model selection strategies in the process. Nevertheless, there remain obstacles to fully appreciate the effect of complex sample characteristics on DE gene detection and the value of improving upon current approaches.

In this paper, we proposed a novel concept for high throughput data analysis involving a heterogeneous sample, i.e. a multi-model handling is intrinsically needed. We presented the theoretical framework that explains why basing inferences on a single model could be problematic in observational microarray studies. The problem arises from the inadequacy of using a single model to describe the complex expression pattern of genes among a heterogeneous sample, which can result in increased number of false discoveries due to bias when a simple model is used or increased random error due to reduced efficiency when a complex model is used. Such effects of model misspecification are hard to avoid because of the existence of genes being affected by different sets of sample characteristics and/or their interactions. We showed through simulation that the single model approaches have inferior performance in DE gene finding in comparison with a multi-model approach should we know the right model for the right set of genes. The magnitude of effects on false discovery depends on the study design, specific biological system and the mechanism underlying expression variation.

We proposed to use BMA approach to improve our ability to identify DE genes. This approach utilizes the Zellner-Siow prior for model parameters. The consistency property of this prior is important as it allows for obtaining a consistent estimate of the distribution of the genes in the model space using Bayes factors. Another choice could be the hyper-g/n prior proposed in Liang et al. (2008). We proposed to use

an iterative procedure to obtain the prior model probabilities so that the estimated distribution of the genes among the model space based on posterior model probabilities matches the estimate based on the Bayes factors. These prior choices allow the efficient computation of the Bayes factors and the posterior inclusion probabilities that does not rely on a MCMC simulation. Our simulation study demonstrated that this approach performed almost as well as the gold standard multi-model approach with true models and better than the single model approaches in gene ranking. The ranking performance was relatively insensitive to a wide range of choice for prior model probabilities. However, accuracy of the FDR directly estimated from the posterior model/inclusion probabilities were sensitive to the prior choice. Our simulation study showed that the proposed empirical prior model probability allowed for reasonably good calibration of posterior model/inclusion probabilities for multiplicity and the estimated FDR was close to the true FDR in settings with moderate to large sample size. In the rare case of a small study with a heterogeneous sample, care needs to be taken when using the empirical prior because the small sample size property of the Zellner-Siow prior is less certain. Nevertheless, it should be pointed out that, multiplicity control in the Bayesian modeling framework remains a challenging and active research area. Further studies on the theoretical aspects of the prior choice for multiplicity control across the multiple genes and multiple models are needed. The current BMA approach is developed under the assumption that relevant covariates are recorded. Should unknown confounders exist, it is possible to extend this method by including the surrogate variables (Leek and Storey, 2007) into the model space. However, we would suggest care in directly incorporating the surrogate variables because these variables were constructed based on residuals from a single model fit of the data.

Finally, through the application of the BMA approach to an observational mircoarray study with unbalanced study design and one with balanced study design, we showed that complex expression patterns did exist when study samples were complex. Previous research has demonstrated the complexities of underlying biological mechanisms for gene expression variation. Genes affected by several common factors, such as age (Tan et al., 2008), gender (Tan et al., 2008; Yang et al., 2006; Delongchamp et al., 2005), smoking (Spira et al., 2004), and drinking alcohol (Lewohl et al., 2001), have been found in different tissue samples. Our study showed that such complexity interfered with the DE gene detection. Notably, the BMA approach was able to avoid missing important genes whose expression patterns were not adequately captured by a single model approach. As an added value, the BMA approach is found to be a flexible tool that allows for more comprehensive characterization of the association between gene expression and the characteristics of the subjects from whom the samples were obtained. And all these can be done within a unified and coherent framework.

Funding

This work was supported by a pilot award to X.K.Z. from the Clinical and Translation Science Center at Weill Cornell Medical College through National Institute of Health (UL1-RR024996), the University of Missouri Research Board Award to F.L., and the funding to A.J.D. from the Flight Attendants Medical Research Institute.

Acknowledgments

The authors thank Drs. Jaya Satagopan and Li-Xuan Qin at Memorial Sloan-Kettering Cancer Center for helpful discussions. *Conflict of Interest:* None declared.

References

BOYLE, J. O., GUMUS, Z. H., KACKER, A., CHOKSI, V. L., JENNIFER, M. B., ZHOU, X. K., ANTE'S, R. K., HUGHES, D., DU, B., JUDSON, B. L., SUBBARAMAIAH, K., AND DANNENBERG, A. J. (2010). Effects of cigarette smoke on the human oral mucosal transcritpome. *Cancer Prevention Reseach*, **3**(3), 266–278.

CAO, J., XIE, X., ZHANG, S., WHITEHURST, A., WHITE, M. A. (2009). Bayesian optimal discovery procedure for simultaneous significance testing. *BMC bioinformatics*, **10**, 5.

CAO, J., ZHANG, S. (2010). Measuring statistical significance for full Bayesian methods in microarray analyses. *Bayesian Analysis*, **2**, 413-428.

CAROLAN, B. J., HARVEY, B.-G., DE, BISHNU P., VANNI, H., AND CRYSTAL, R. G. (2008). Decreased expression of Intelectin 1 in the human airway epithelium of smokers compared to nonsmokers. *Journal of Immunology*, **181**, 5760–67.

CONLON, E. M., SONG, J. J., AND LIU, J. S. (2006). Bayesian models for pooling microarray studies with multiple sources of replications. *BMC Bioinformatics*, **7**, 247.

DELONGCHAMP, R. R., VELASCO, C., DIAL, S., AND HARRIS, A. J. (2005). Genome-wide estimation of gender differences in the gene expression of human livers: statistical design and analysis. *BMC bioinformatics*, **6 Suppl 2**, S13.

EFRON, B. (2008). Microarrays, empirical Bayes and the two-groups model. *Statistical Science*. **23**(1):1–22.

HELLER, R., MANDUCHI, E., & SMALL, D. S. (2009). Matching methods for observational microarray studies. *Bioinformatics*, **25**(7), 904–909.

HOETING, J. A., MADIGAN, D., RAFTERY, A. E., AND VOLINSKY, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical Science*, **14**(4), 382–401.

HUMMEL, M., MEISTER, R., & MANSMANN, U. (2008). GlobalANCOVA: exploration and assessment of gene group effects. *Bioinformatics*, **24**, 78–85.

ISHWARAN, H., AND RAO, J. S. (2003). Detecting Differentially Expressed Genes in Microarrays Using Bayesian Model Selection. *Journal of the American Statistical Association*. **98**(462), 438–455

JEFFERY, I., HIGGINS, D., AND CULHANE, A. (2006). Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data. *BMC Bioinformatics*, **7**, 359.

KASS, R. E., RAFTERY, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, **90**(430):773–795.

LEEK, J. T., STOREY, J. D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics*, **3**(9):1724–1735.

LEWOHL, J. M., DODD, P. R., MAYFIELD, R. D., AND HARRIS, R. A. (2001). Application of DNA microarrays to study human alcoholism. *Journal of biomedical science*, **8**, 28–36.

LIANG, F., PAULO, R., MOLINA, G., CLYDE, M. A., AND BERGER, J. O. (2008). Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association*, **103**, 410–423.

MÜLLER, P., PARMIGIANI, G., RICE, K. (2007) FDR and Bayesian multiple comparisons rules. In *Bayesian Statistics 8*(Bernardo, J. M., Bayarri, M., Berger, J., et al., eds.). Oxford: Oxford University Press.

NEWTON, M., NOUEIRY, A., SARKAR, D., AND AHLQUIST, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics*, **5**(2), 155–176.

POTTER, J. D. (2003). Epidemiology, cancer genetics and microarrays: making correct inferences, using appropriate designs. *Trends in genetics : TIG*, **19**, 690–695.

RAO, P. (1971). Some notes on misspecification in multiple regressions. *The American Statistician*, **25**, 37–39.

RAO, P. (1973). Some notes on the errors-in-variables model. *The American Statistician*, **27**, 217–218.

ROSENBERG, S. H., AND LEVY, P. S. (1972). Note: a characterization on misspecification in the general linear regression model. *Biometrics*, **28**, 1129–1133.

SARTOR, M. A., TOMLINSON, C. R., WESSELKAMPER, S. C., SIVAGANESAN, S., LEIKAUF G. D., MEDVEDOVIC, M. (2006). Intensity-based hierarchical Bayes method improves testing for differentially expressed genes in microarray experiments. *BMC bioinformatics*, **7**, 538.

SEBASTIANI, P., XIE, H., AND RAMONI, M. F. (2006). Bayesian Analysis of Comparative Microarray Experiments by Model Averaging. *Bayesian Analysis*, **1**(4), 707–732.

SCHEID, S., & SPANG, R. (2007). Compensating for unknown confounders in microarray data analysis using filtered permutations. *Journal of Computational Biology*, **14**, 669–681.

SMYTH, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology*, **3**, Article3.

SPIRA, A., BEANE, J., SHAH, V., LIU, G., SCHEMBRI, F., YANG, X., PALMA, J., AND BRODY, J. S. (2004). Effects of cigarette smoke on the human airway epithelial cell transcriptome. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 10143–10148.

STOREY, J. D (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **64**(3), 479-498.

STOREY, J. D., TIBSHIRANI, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America*. **100**(16), 9440–9445.

TAN, Q., ZHAO, J., LI, S., CHRISTIANSEN, L., KRUSE, T. A., AND CHRISTENSEN, K. (2008). Differential and correlation analyses of microarray gene expression data in the CEPH Utah families. *Genomics*, **92**, 94–100.

TROESTER, M. A., MILLIKAN, R. C., AND PEROU, C. M. (2009). Microarrays and epidemiology: ensuring the impact and accessibility of research findings. *Cancer epidemiology, biomarkers & prevention*, **18**, 1–4.

WEBB, P. M., MERRITT, M. A., BOYLE, G. M., AND GREEN, A. C. (2007). Microarrays and epidemiology: not the beginning of the end but the end of the beginning... *Cancer epidemiology, biomarkers & prevention*, **16**, 637–638.

WU, X. L., GIANOLA, D., ROSA, G. J.M., WEIGEL, K. A. (2010). Bayesian model averaging for evaluation of candidate gene effects. *Genetica*. **138**(3), 395–407.

XU, L., CRAIU, R. V., AND SUN, L. (2011). Bayesian methods to overcome the winners curse in genetic studies. *The Annals of Applied Statistics*. **5**(1), 201–231.

YANG, X., SCHADT, E. E., WANG, S., WANG, H., ARNOLD, A. P., INGRAM-DRAKE, L., DRAKE, T. A., AND LUSIS, A. J. (2006). Tissue-specific expression and regulation of sexually dimorphic genes in mice. *Genome Research*, **16**, 995–1004.

YEUNG,K. Y., BUMGARNER, R. E., AND RAFTERY,A. E. (2005). Bayesian Model Averaging: Development of an improved multi-class, gene selection and classification tool for microarray data. *Bioinformatics*, **21**(10):2394–2402.

ZELLNER, A., AND SIOW, A. (1980). Posterior odds ratios for selected regression hypotheses. In*Bayesian Statistics: Proceedings of the First International Meeting Held in Valencia (Spain)*(Bernardo, J. M., DeGroot, M. H., Lindley D. V., and Smith,A.F.M. eds.) 585–603. Valencia: Valencia University Press.